软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007708] [CSTR: 32375.14.jos.007708]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于像素级-块级伪影协同建模的 AI 生成图像模型溯源*

范宇祺[1], 陈嘉欣[2], 廖　鑫[1], 陈昌盛[3], 章登勇[2]

[1](湖南大学 网络空间安全学院, 湖南 长沙 410082)

[2](长沙理工大学 计算机学院, 湖南 长沙 410076)

[3](深圳北理莫斯科大学 工程学院, 广东 深圳 518172)

通信作者: 陈嘉欣, E-mail: jxchen@csust.edu.cn; 廖鑫, E-mail: xinliao@hnu.edu.cn

摘　要: 随着扩散模型和生成对抗网络的快速发展, AI 生成图像在内容创作中的应用日益广泛, 同时也对内容可信性与生成来源溯源提出新的挑战. 相较于仅区分图像真伪, 准确判定生成图像的来源模型具有更高的实际应用价值, 例如在生成内容平台监管、恶意伪造内容追溯以及生成模型责任追踪等场景中, 能够为内容来源识别与可信治理提供重要技术支撑. 不同生成模型在合成过程中往往会在像素层面留下细粒度高频伪影, 同时在结构与全局层面形成具有模型特征的生成模式一致性差异. 这种跨尺度的伪影现象表明, 现有的仅依赖单一尺度或单一语义层面的特征表示方法, 难以同时刻画细粒度伪影与全局生成风格差异, 从而限制在复杂多模型场景下归属性能. 针对上述不足, 提出一种像素到块级伪影协同建模 (collaborative pixel-block artifact modeling, CPB-AM) 框架, 用于 AI 生成图像的模型溯源任务. 该方法采用双分支结构对不同层级的生成伪影进行协同建模, 像素级分支通过高频增强操作提取局部伪造痕迹, 并学习与生成域相关的细粒度伪影特征; 块级分支将图像划分为局部块, 并通过频率感知的注意力聚合机制对块间的相关性进行建模. 该注意力机制通过对低频平滑成分进行自适应聚合, 显式分离结构一致性表示与残差成分, 从而更有效地刻画生成模型在空间结构与整体风格层面的差异. 同时, 在跨尺度融合方面, 设计跨分支注意力引导融合模块, 是一种引导式特征融合机制, 通过构建像素级伪影特征对块级表示的调制权重, 实现细粒度残差信息对结构建模过程的引导与约束, 从而增强跨尺度伪影信息的互补性. 在包含真实图像与多种生成模型的数据集上的实验结果表明, 所提方法在闭集生成模型溯源任务中能够取得优于现有方法的分类性能; 在未见生成模型参与测试的开集实验设置下, 该方法在未知模型拒识与归属判别方面同样表现出较强的鲁棒性. 实验结果验证了像素到块级伪影协同建模策略在提升生成模型溯源准确性与开放场景适应能力方面的有效性.

关键词: AI 生成图像溯源; 像素-块级伪影协同建模; 引导式特征融合; 开放集拒识

中图法分类号: TP18

中文引用格式: 范宇祺, 陈嘉欣, 廖鑫, 陈昌盛, 章登勇. 基于像素级-块级伪影协同建模的AI生成图像模型溯源. 软件学报. http://www.jos.org.cn/1000-9825/7708.htm

英文引用格式: Fan YQ, Chen JX, Liao X, Chen CS, Zhang DY. Collaborative Pixel-block Artifact Modeling for AI-generated Image Model Source Attribution. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7708.htm

* 基金项目: 国家自然科学基金区域创新发展联合基金重点项目 (U22A2030); 国家自然科学基金青年基金 (62402062); 国家自然科学基金面上项目 (62371301); 国家重点研发计划 (2024YFF0618800); 湖南省自然科学基金青年基金 (2025JJ60415); 湖南省自然科学基金杰出青年基金 (2024JJ2025); 湖南省重点研发计划 (2024AQ2027, 2025AQ2022); 湖南省重大科技攻关 (2025QK2008); 湖南省教育厅优秀青年项目 (25B0221); 岳麓山工业创新中心项目 (2025YCII0222)
收稿时间: 2026-03-25; 修改时间: 2026-04-16; 采用时间: 2026-05-21

Collaborative Pixel-block Artifact Modeling for AI-generated Image Model Source Attribution

FAN Yu-Qi[1], CHEN Jia-Xin[2], LIAO Xin[1], CHEN Chang-Sheng[3], ZHANG Deng-Yong[2]

[1](College of Cyber Science and Technology, Hunan University, Changsha 410082, China)

[2](School of Computer Science and Technology, Changsha University of Science and Technology, Changsha 410076, China)

[3](Faculty of Engineering, Shenzhen MSU-BIT University, Shenzhen 518172, China)

Abstract: With the rapid development of diffusion models and generative adversarial networks, AI-generated images are increasingly used in content creation, while also posing new challenges to content trustworthiness and source attribution. Compared with merely distinguishing authentic images from fake ones, accurately identifying the source generative model of generated images has greater practical significance. For instance, this capability can provide important technical support for content source identification and trustworthy governance in scenarios such as generated content platform regulation, tracing of maliciously forged content, and responsibility tracking of generative models. During the synthesis process, different generative models often leave fine-grained high-frequency artifacts at the pixel level, while also producing model-specific differences in generative pattern consistency at the structural and global levels. This cross-scale artifact phenomenon indicates that existing methods relying solely on single-scale or single-semantic-level feature representations are insufficient for simultaneously capturing fine-grained artifacts and global generative style differences, thus limiting attribution performance in complex multi-model scenarios. To address these issues, this study proposes a collaborative pixel-block artifact modeling framework (CPB-AM) for AI-generated image model source attribution. The proposed method adopts a dual-branch architecture to jointly model generative artifacts at different levels. The pixel-level branch extracts local forgery traces through high-frequency enhancement operations and learns fine-grained artifact features associated with the generative domain. The block-level branch partitions the image into local blocks and models inter-block correlations through a frequency-aware attention aggregation mechanism. This attention mechanism adaptively aggregates low-frequency smooth components and explicitly separates structural consistency representations from residual components, enabling more effective characterization of differences among generative models in terms of spatial structure and overall style. Furthermore, for cross-scale fusion, a cross-branch attention-guided fusion module is designed as a guided feature fusion mechanism. By constructing modulation weights from pixel-level artifact features for block-level representations, the module enables fine-grained residual information to guide and constrain the structural modeling process, thus enhancing the complementarity of cross-scale artifact information. Experimental results on datasets containing real images and multiple generative models demonstrate that the proposed method achieves superior classification performance compared with existing approaches in closed-set generative model source attribution tasks. Under open-set experimental settings involving unseen generative models, the proposed method also exhibits strong robustness in both unknown model rejection and attribution discrimination. These results verify the effectiveness of the collaborative pixel-block artifact modeling strategy in improving generative model source attribution accuracy and adaptability to open scenarios.

Key words: AI-generated image attribution; collaborative pixel-block artifact modeling; guided feature fusion; open-set rejection

近年来, 随着生成对抗网络 (generative adversarial network, GAN) 和扩散模型 (diffusion model) 等生成模型的快速发展, AI 生成图像在视觉质量、多样性和真实感等方面取得了突破性进展, 并已广泛应用于内容创作、数字艺术、虚拟现实以及多媒体生产等实际场景[1,2]. 这类技术在显著提升创作效率与表现力的同时, 也在一定程度上模糊了真实图像与合成图像之间的界限, 使得普通用户甚至专业系统难以仅凭直观感知对其真实性与来源进行判断[3]. 然而, 伴随着 AI 生成图像在网络空间中的大规模传播, 其在内容可信性、版权保护、舆情操纵和数字取证等方面所引发的潜在风险日益凸显. 例如, 未经标注的生成图像可能被用于伪造新闻、误导公众认知, 或在版权纠纷中造成责任归属不清的问题. 在这样的背景下, 如何对 AI 生成图像进行有效分析, 并进一步准确追溯其生成来源模型, 已成为多媒体取证、内容安全与人工智能治理领域中亟需解决的关键问题之一[4].

AI 生成图像模型溯源旨在判定图像的具体生成模型, 这一任务不仅具有重要的应用价值, 能够为平台监管、责任追溯和生成模型滥用分析提供证据支持, 同时在学术研究中也有助于揭示不同生成模型在图像空间中的差异性. 现有的 AI 生成图像检测任务多集中于真假图像的判别, 例如, Karageorgiou 等人[5]使用自监督学习训练频谱重建模型, 并基于频谱重建相似度捕捉生成图像; Tao 等人[6]提出的 SAGNet 方法引入对抗训练机制, 使特征提取器学习到难以被类别判别器捕获的表示, 从而抑制图像语义信息对判别结果的干扰, 突出模型生成过程中产生的语义无关伪影特征. 这些方法虽然有效实现了真假图像的鉴定, 但它们的任务范围仅限于二分类问题, 大多无法有效地进行生成模型的溯源. 生成模型溯源面临更大的挑战: 一方面, 生成模型的“模型指纹”往往在图像中留下的是

低对比度的伪影特征, 这些特征在后处理、压缩或跨域传输后容易被削弱或消失; 另一方面, 生成模型的“模型指纹”涉及多个层次, 如像素统计特性、局部纹理结构和高层语义一致性等, 这使得单一尺度的特征表示方法难以全面刻画图像的生成来源.

针对上述问题, 已有研究尝试从深层表示特征、人工指纹特征等相关特征提取学习[4,7]、重建误差分析[8,9]或图文结合语义表示[10,11]等多个角度对生成模型进行区分. 在特征提取与判别学习方向, Yu 等人[7]提出通过在训练阶段向生成模型中注入可学习的人工指纹, 使生成图像在频域或像素统计层面嵌入稳定且可追踪的模式, 从而实现可验证的来源归因. 在重建误差分析与潜空间建模方向, Wang 等人[9]提出通过构建反向映射网络, 将生成图像映射回潜空间, 并利用潜变量重建一致性作为归因依据, 从而在无需人工水印的情况下实现来源追踪. 在图文联合表示与语义一致性建模方面, Sha 等人[10]提出结合视觉特征与文本提示信息, 通过跨模态对齐机制分析图像-文本一致性差异, 实现生成图像检测与归因联合建模. 尽管上述方法在生成模型归因方面取得了一定进展, 但仍然存在一些局限性. 一方面, 多数方法主要依赖单一类型的特征表征进行判别, 或仅利用深层语义特征进行归因分析, 难以同时充分刻画图像中的局部纹理细节与整体结构模式; 另一方面, 在多模型混合的复杂场景下, 这类方法往往仍然面临泛化能力不足的问题. 此外, 在面对未见生成模型时, 部分方法缺乏有效的区分能力, 难以满足开放环境下的归因拒识需求.

基于上述观察, 本文提出一种基于像素级与块级特征协同建模的生成模型归因框架. 其核心思想在于: 生成模型在图像中留下的指纹信息在多个尺度具有不同的特征表现, 不同尺度的特征能够反映生成过程中不同类型的伪影信息. 已有研究表明, 生成模型在像素统计与高频残差层面往往会留下可被捕捉的特征模式[12], 如高频残差和局部纹理扰动的像素级特征, 这有助于刻画生成过程中引入的底层统计伪影[13], 而如基于 Transformer 表示的块级特征则能够建模图像中不同区域的结构性模式及语义相关的全局依赖关系[14]. 经过初步分析, 可以看出二者在判别能力与鲁棒性方面具有互补性.

具体而言, 首先设计了一个像素级特征分支, 该分支进行像素级高频建模 (pixel-level high-frequency modeling, HF), 利用二阶差分算子对输入图像进行高频增强, 以显式突出生成过程中的细粒度伪影信息; 在骨干网络结构上, 本文基于卷积框架构建多层级特征提取网络, 并在全局池化后引入双分支投影头 (artifact head), 实现域相关特征与域无关特征的表示解耦, 并保留域相关特征. 同时, 构建一个块级特征分支, 引入 ViT (vision Transformer)[15]与基于多头自注意力机制的块级高频伪影建模 (patch-level high-frequency artifact modeling, PHF) 模块, 形成融合 ViT 的块级模式感知分支, 用于对图像进行块级建模, 从而捕捉高维语义残留及跨区域结构特征. 为实现不同层级特征之间的有效协同, 本文进一步引入交叉注意力 (cross-attention) 机制[16,17], 提出一种跨分支注意力引导融合 (cross-branch attention-guided fusion, CBAGF) 机制, 利用像素级细粒度指纹对块级结构特征的选择与聚焦过程进行引导, 实现跨分支信息的对齐与互补融合.

实验结果表明, 所提框架在包含多种主流生成模型的数据集上均表现出稳定且具有竞争力的归因性能. 通过像素级特征的指纹解耦与块级特征的结构建模, 并结合引导式融合机制, 本文方法能够有效保留生成图像中的多尺度判别信息, 显著提升模型在复杂场景下的泛化能力. 此外, 在未见生成模型的开放集评估场景中, 所提框架基于能量评分在统计意义上展现出稳定的未见模型拒识能力, 说明该框架的鲁棒性与泛化潜力, 具备一定的实际应用价值.

总体来说, 本文的主要贡献如下.

(1) 构建像素级-块级协同建模溯源架构: 不同于现有方法主要依赖单尺度特征建模或单一判别空间学习的策略, 本文从生成模型“伪影分布跨层级存在”的特性出发, 设计像素级高频增强与块级结构建模相结合的双分支架构. 在像素层面, 通过高频强化机制显式放大生成过程中残留的微观统计异常, 抑制内容语义对低层伪影建模的干扰; 在块级层面, 通过结构建模捕获宏观纹理与上下文依赖关系. 该设计实现了从细粒度局部指纹到全局结构模式的多尺度覆盖, 有效缓解了单尺度建模易忽略弱伪影或过分关注高层语义信息的问题.

(2) 设计跨分支引导融合机制与域相关解耦 (domain-aware decoupling, DA-Decoupling) 模块: 针对不同生成模型之间存在模型特有指纹特征的特点, 本文通过域相关表示解耦机制从像素分支的高频表征中独立分化出专用的

域相关特征 (domain-specific feature), 用于刻画不同生成模型之间的差异性指纹特征, 该结构能够在统一框架下分离出“模型特有信号”, 避免特征混叠导致的判别边界模糊问题, 从而增强多生成模型归因任务中的特征可分性. 同时, 为强化不同尺度特征之间的协同关系, 本文提出跨分支注意力引导融合模块, 通过门控单元动态调节结构特征与高频残差信号的融合比例, 并利用像素级伪影响应对块级结构建模进行引导, 实现跨层级信息的互补增强. 该机制有助于在保持结构语义稳定性的同时强化细粒度伪影表达, 提高模型对弱差异生成模型的区分能力.

(3) 在多模型混合数据集与跨模型测试设置下, 本文方法在生成模型归因准确率、平均性能指标以及开放场景下的区分能力等方面均优于多种代表性方法. 消融实验进一步验证像素级高频增强、跨分支引导融合以及域相关解耦机制对性能提升的独立贡献, 证明了所提出协同建模策略在多层级伪影刻画与复杂分布泛化方面的有效性.

1 相关工作

1.1 AI 生成图像真伪检测

在生成对抗网络和扩散模型快速发展的背景下, AI 生成图像检测成为数字媒体取证领域的重要研究方向[18]. 该任务的目标是判定输入图像是否由生成模型合成, 即进行真实与伪造的二分类判别. 已有研究表明, 生成图像在合成过程中往往会在频域分布、局部纹理结构以及空间一致性等方面引入异常特征, 这些异常为检测图像真伪提供了重要依据[19].

早期研究主要针对生成对抗网络生成图像的检测问题, 通过构建频域特征等低层信号表征来刻画生成图像在分布层面的偏离特性, 例如利用频谱能量分布等进行判别[20]. 随着深度学习的发展, 基于卷积神经网络的检测方法逐渐成为主流, 这类方法通过自动学习图像的空间结构与纹理模式, 实现对生成图像的有效识别, 并在多种生成模型场景下取得较好性能[21,22]. 近年来, 随着扩散模型的兴起, 相关研究开始关注扩散生成过程中引入的结构不一致性与噪声分布特征, 并探索面向扩散模型的检测方法.

相比 GAN 生成图像, 扩散模型生成结果在视觉质量和统计分布上更接近真实图像, 使得传统检测方法的判别能力下降. 针对这一问题, 一些研究开始从生成机制角度进行建模. 例如 LaRE^2 方法[21]通过分析扩散模型的潜空间重建误差, 将输入图像映射至潜空间并计算重建偏差, 以此作为判别依据, 从而实现对扩散生成图像的有效检测. 这类方法表明, 从生成过程出发刻画图像分布偏离, 有助于提升检测模型的鲁棒性. 另一方面, 部分研究从判别特征构造角度出发, 强调对真实与生成图像之间差异性的显式建模, 比如 Yang 等人[22]通过学习真实图像与生成图像之间的特征不一致性, 构建统一的判别表示, 从而提升模型在大规模与跨分布场景下的泛化能力. 该类方法表明显式建模图像中的差异模式有助于捕获生成过程中引入的潜在伪影特征.

1.2 AI 生成图像模型溯源

在 AI 生成图像检测任务取得一定进展的基础上, 针对 AI 生成图像的来源归属问题逐渐受到关注[23]. 生成图像模型溯源旨在判定给定图像由何种生成模型产生, 相比仅进行真伪判别, 该任务在内容监管、版权保护及虚假信息治理等场景中具有更高的实际价值. 早期研究主要聚焦于 GAN 生成图像, 通过挖掘生成过程中引入的统计偏差或伪影特征, 实现对不同生成模型的区分. Bui 等人[4]提出的 RepMix 方法在卷积神经网络中间层对来自不同图像的特征进行随机比例混合, 并训练模型预测混合比例, 从而学习生成模型特有的细微伪影特征. 在此基础上, 部分研究开始关注更加复杂的应用场景, 如未见模型或样本分布变化. Yang 等人[23]提出渐进式开放空间扩展 (progressive open space expansion, POSE) 框架, 通过逐步扩展闭集样本周围的特征空间来模拟潜在未见模型的指纹分布, 实现闭集模型归属与开集拒识的统一建模. 此外, Girish 等人[24]将特征学习、OOD 检测与聚类过程整合为统一流程, 实现了开放世界场景下 GAN 生成图像的发现与归属.

随着文本生成图像 (text-to-image, T2I) 扩散模型[25]的广泛应用, 研究重点逐渐从 GAN 扩展至扩散模型及其变体. Sha 等人[10]进一步结合图像单模态与图文混合信息, 在跨数据集设置下验证了方法的泛化能力. 与此同时, Wang 等人[9]的方法利用扩散模型自身的编码器与重建误差, 实现无需人工水印的高精度模型归属, 通过这种方式尝试摆脱对显式监督分类器的依赖, 从重建或反演角度进行生成模型溯源, 在此方面, ReTD 方法[26]同样采用重建

的思想, 通过重建放大不同生成模型在图像生成过程中的细微差异, 再对这些差异进行建模以实现来源模型判别.

1.3 多尺度与跨层级特征建模

在数字图像取证 (digital image forensics) 研究中, 特征建模方式对伪造检测等取证任务具有重要影响. 已有的深度伪造检测研究发现, 伪造图像在生成或篡改过程中往往会引入多种异常痕迹, 包括频域能量分布异常、图像融合边界伪影以及局部纹理结构失真等[27–29], 例如 F3-Net[27]通过挖掘频域信息中的异常模式实现人脸伪造检测, 而 Face X-ray 方法[28]则通过建模图像融合区域的伪影特征识别深度伪造内容. 除此之外, 也有研究从更一般的视觉表征角度出发, 探索能够跨不同伪造方法共享的判别特征, 例如 UCF 方法[30]通过挖掘不同深度伪造模型之间的共性特征来提升检测模型的泛化能力. 这些研究利用深度神经网络从空间结构特征、频域信息以及局部纹理模式等多个角度学习伪造图像的判别特征, 从而实现对伪造内容的检测.

在特征建模层面, 现有 AI 生成图像溯源方法也呈现出明显的研究侧重点差异. 有些方法主要关注像素级或频域层面的低层统计特征, 通过捕获生成过程引入的局部伪影来实现模型区分[4,13,31]. 另一类方法则依赖深度网络提取的高层语义表示, 刻画图像整体结构或生成风格特征[11,32], 以 Lee 等人[32]为代表的工作是基于 CLIP-ViT 视觉编码器提取高层语义嵌入, 并构建可学习表示以支持小样本与类增量场景下的生成模型归因. 其核心思想是利用预训练视觉语言模型的强语义表达能力, 将不同生成模型的整体风格差异映射到统一的高维语义空间中, 并通过增量学习策略扩展模型类别. 该方法在类增量与少样本归因任务中表现出一定优势, 表明高层语义特征能够有效刻画生成模型的宏观表达模式. 然而, 由于其表示主要依赖语义嵌入, 当图像内容分布变化或存在跨域测试时, 语义信息可能掩盖生成机制的细粒度差异, 从而影响归因稳定性. 另一方面, Xu 等人[11]针对文本到图像扩散模型的来源归因问题, 基于深度视觉表征提取不同扩散模型生成图像的全局结构特征, 并构建分类器实现多模型区分. 该方法强调扩散模型在整体结构组织与风格表达上的差异, 在闭集设置中取得了较好的性能.

近年来, 也有部分研究开始尝试引入多层级或多尺度信息以增强特征判别能力. 例如, Koutlis 等人[31]利用编码器中间层特征进行合成图像检测, 验证了中低层表示在捕获生成伪影方面的有效性, 而 ManiFPT 方法[33]从指纹建模角度分析了不同生成模型在特征空间中的分布特性.

1.4 开集评估与未见模型鲁棒性

在实际应用中, 生成模型的种类和版本持续改进优化, 使得模型溯源系统不可避免地需要面对未见或未知生成源. 为此, 部分研究将模型归属问题扩展至开集或开放世界设置. 有些方法通过模拟不同生成模型的指纹分布, 探索零样本或开集条件下的模型归属能力[23,34]. 在此基础上, 一些研究也尝试对未见生成模型进行聚类分析[35]. 为了尝试解决开放场景下溯源鲁棒性的问题, Sun 等人[36]提出的框架通过结合全局和局部特征对齐与置信度引导的软伪标签策略, 在半监督学习框架下充分利用未标注数据, 从而提升开放世界场景下深度伪造生成方法归属的泛化能力和未知类别发现能力.

在评估层面, 除传统闭集分类准确率外, 越来越多的研究引入基于能量、置信度或距离的拒识评分, 并使用 AUC、OSCR 等指标来衡量模型在未见生成源条件下的区分能力[23]. 这些研究表明, 引入开集评估能够更加全面地刻画模型溯源方法的泛化性能, 也为方法设计提供了更为严格和贴近实际应用的验证标准.

2 基于像素到块级伪影协同建模的 AI 生成图像模型溯源框架

2.1 CPB-AM 基本框架

本文提出了一种基于像素到块级伪影协同建模 (collaborative pixel-block artifact modeling, CPB-AM) 的生成图像模型溯源框架, 其整体结构如图 1 所示. 该框架由两个主要分支组成.

● 细粒度像素级高频伪影建模分支: 该分支以输入图像 X 为起点, 首先提取像素层面的高频特征, 采用基于卷积框架的多层级特征提取网络. 该网络通过逐层提取不同尺度的特征, 能够高效地捕捉图像中的细粒度伪影信息, 在全局池化后引入域相关投影头用于显式分离域相关特征, 使其刻画不同生成模型之间的差异性指纹特征, 便于后续生成模型溯源与拒识.

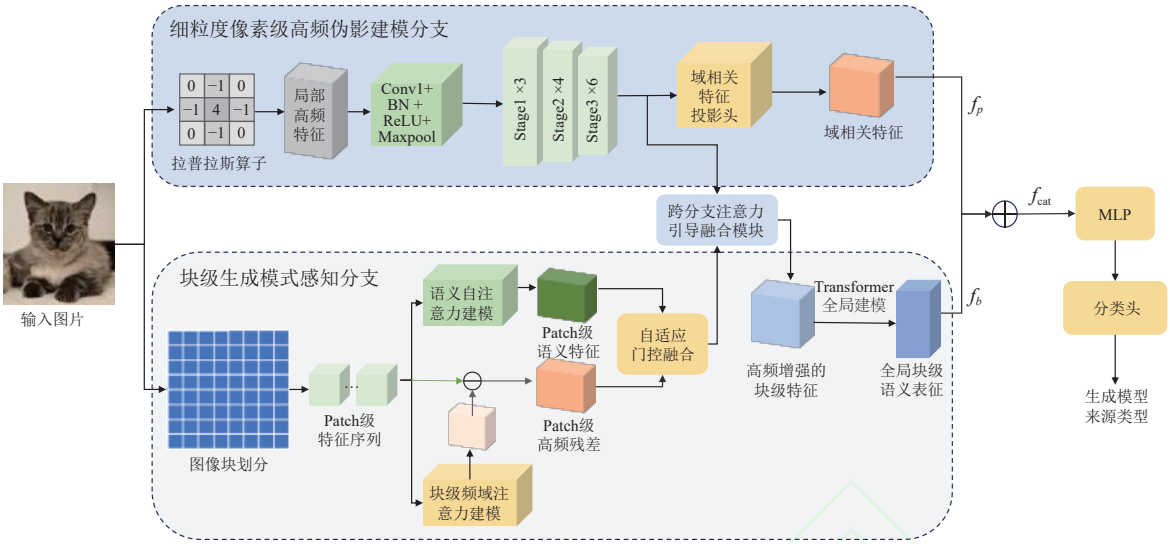


图 1 基于像素级-块级伪影协同建模的生成图像模型溯源基本框架图

● 块级生成模式感知分支: 该分支从图像块级别出发, 采用融合 ViT 网络, 并结合多头自注意力 (multi-head self-attention, MHSA) 机制[16]对不同图像块之间的全局依赖关系进行建模, 从而捕获图像的结构组织模式与跨区域上下文关系. 同时, 通过基于注意力的残差建模方式, 进一步提取块级高频伪影信息, 以刻画生成模型在区域结构与整体模式上的差异.

为了充分利用不同层级特征的互补性, 本文设计了一种基于交叉注意力的跨分支注意力引导融合模块, 使得像素级高频伪影特征能够对块级特征的选择进行一定程度的引导. 最终, 将来自两个分支的特征联合用于生成模型归因与开放场景模型拒识, 实现了从微观纹理到全局结构的多尺度表征能力.

2.2 像素级与块级高频建模差异分析

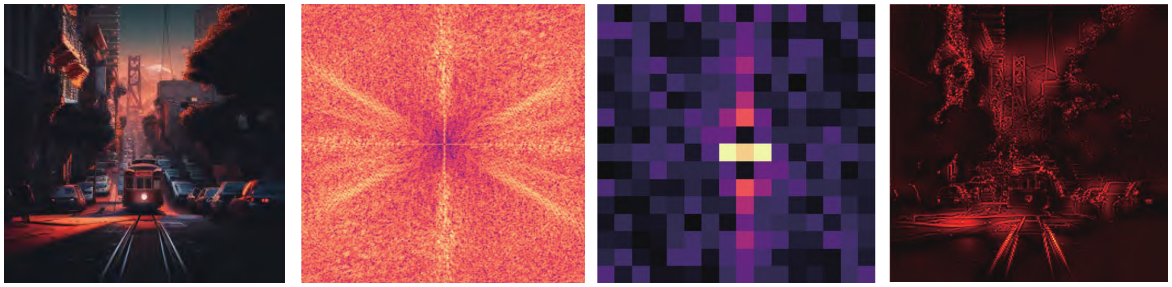
在具体介绍 CPB-AM 框架中像素级与块级伪影建模分支之前, 本文首先从理论与统计层面对不同高频建模方式的表征差异进行分析, 以验证跨层级协同建模的必要性.

从信号处理角度来看, 拉普拉斯 (Laplacian) 算子是一种二阶微分算子, 其响应主要集中在图像灰度变化剧烈的位置, 强调局部边缘与纹理突变区域[37]. 在本文提出的像素级高频建模分支中, 通过引入拉普拉斯算子的高频残差提取操作, 输入图像中的细粒度伪影信息得到了强化, 从而突出了生成过程中残留的局部统计异常与纹理扰动; 而块级高频建模通常通过卷积网络在较大感受野下学习残差信息, 其响应不仅受局部梯度变化影响, 还会融合上下文结构与区域级模式信息. 因此, 两种高频表示在理论上分别对应局部微分响应与上下文残差建模两类不同的频域表达机制[38].

为直观验证上述分析, 图 2 给出了两种高频建模方式的频谱与空间响应对比结果. 图 2(a) 为输入的原始图像; 图 2(b) 展示了通过拉普拉斯算子提取的像素级高频响应在频域中的幅值分布, 其能量主要集中于高频边缘区域; 图 2(c) 为块级分支学习到的高频特征在频域中的响应分布, 可以观察到其频谱分布更为分散, 并在中高频区域呈现结构性聚集趋势, 反映出块级建模对区域级模式的整合能力; 图 2(d) 进一步给出两种高频特征在空间域中的差异热力图, 可以观察到差异响应主要集中在物体边缘及细粒度纹理区域, 说明像素级高频建模在这些局部结构变化剧烈的位置产生了更强响应, 而块级特征在这些区域的响应相对平滑, 从而使得两者在空间分布上呈现差异.

为了避免单一样本分析带来的偶然性影响, 本文在包含 1 000 张不同生成模型图像的数据子集上, 对像素级拉普拉斯高频表示与块级高频特征进行余弦相似度统计分析. 设像素级高频特征为 F_l, 块级高频特征为 F_b, 则其余弦相似度定义为:

$$\cos(\theta) = \frac{F_l \cdot F_b}{\|F_l\| \|F_b\|} \tag{1}$$



(a) 原始图像 (b) 像素级高频频谱图 (c) 块级分支高频频谱图 (d) 高频分布差异热力图

图 2 像素级与块级高频特征的空间与频域对比可视化

统计结果表明, 在 1 000 张图像样本上, 两类高频表示的平均余弦相似度为 0.036. 该结果说明两种高频特征在特征空间中呈现出显著低相关性, 反映出其关注区域与表达模式存在本质差异. 结合理论分析与统计结果可以发现, 像素级高频建模更敏感于局部微观扰动, 而块级高频表示则更倾向于整合区域级结构残差信息. 二者在频域分布、空间响应模式及特征相关性方面均表现出明显差异. 因此, 仅依赖单一尺度高频建模难以全面刻画生成模型在不同层级上产生的伪影特征.

由于两类特征在特征空间中呈现出显著低相关性, 这种差异性为互补信息提供了基础. 所以在此基础上, 本文进一步通过结构化融合机制将上述差异性转化为互补特征. 具体而言, 像素级分支提取的高频特征主要刻画生成过程中的局部统计异常, 具有较强的内容不变性, 可作为稳定的模型指纹; 而块级分支则通过自注意力机制建模跨区域的结构关系, 提供语义上下文与空间一致性约束. 在跨分支融合过程中, 本文采用单向引导策略, 将像素级高频特征作为先验信息注入块级分支, 通过注意力机制对其进行自适应加权, 从而在全局结构建模过程中选择性地强化具有判别性的局部伪影模式.

从机制上看, 该过程并非简单的特征拼接或叠加, 而是使用协同建模方式实现多尺度信息融合, 具体而言, 像素级特征提供语义无关的稳定判别线索, 块级特征则根据全局结构对这些线索进行上下文约束与空间对齐, 使得局部伪影能够在不同语义场景下保持一致表达. 同时, 结合单向梯度控制机制, 可以避免高层语义信息对低层统计特征的反向干扰, 从而保证指纹特征的稳定性.

因此, 多尺度协同建模不仅在特征空间上实现了低相关性信息的联合编码, 更通过结构化融合机制增强了不同层级特征之间的协同作用, 使模型能够同时捕获细粒度纹理异常与结构级偏差信息. 这种局部指纹和全局结构的互补表征方式能够有效缓解语义变化带来的干扰, 提升生成模型溯源任务中的判别能力与泛化性能.

2.3 细粒度像素级高频伪影建模分支

该分支旨在从像素和局部纹理层面显式建模 AI 生成图像中由生成模型与采样机制引入的细粒度伪影特征. 不同于仅依赖高层语义信息的判别方式, 本分支从高频信号出发, 通过深层卷积网络和用投影伪影表征头, 学习生成模型相关的判别性表示, 为后续的模型溯源和开集拒识提供稳定的底层指纹支撑.

从生成机制角度来看, GAN 与扩散模型等主流生成模型在图像合成过程中依赖卷积运算、上采样 (如反卷积或插值) 以及逐步去噪等操作. 这些过程在引入真实感的同时, 也不可避免地产生了具有结构规律性的偏差, 例如由上采样操作引入的周期性纹理、由卷积核共享带来的局部统计一致性偏差, 以及由采样策略导致的频谱分布不均衡[39]. 这类偏差主要体现在图像的细粒度纹理层面, 在频域中通常表现为高频分量的异常增强或分布失衡[12].

进一步地, 从语义与频率的关系来看, 图像的低频成分主要承载整体结构与语义信息, 而高频成分则更多反映局部纹理细节与生成过程中的误差累积. 因此, 高频伪影的形成主要由生成模型结构与采样机制决定, 而非具体的图像内容. 这意味着, 在不同语义类别与场景下, 只要生成模型保持一致, 其所引入的高频统计模式往往就具有较

强的一致性, 从而表现出相对稳定的分布特性. 相较之下, 高层语义特征则随内容变化显著, 难以作为稳定的模型指纹[7].

基于上述分析, 本文通过拉普拉斯算子对输入图像进行高频增强, 显式提取高频信号, 以强化对局部纹理异常的响应. 在此基础上, 利用深层卷积网络对高频残差进行建模, 在局部感受野内统计其分布模式, 并通过投影头将其映射至判别性特征空间. 该过程本质上实现了从生成结构偏差到可学习特征表示的映射, 使得模型能够捕获具有跨内容一致性的伪影模式, 从而提升生成来源判别的稳定性与泛化能力.

本文首先将输入的任意分辨率的三通道 RGB 图像 X, 统一裁剪至 224×224 大小的 RGB 图像, 以适应模型输入. 针对生成图像伪影在频域中表现出的高频分量, 为了聚焦这些图像的微观痕迹, 对裁剪后的图片进行拉普拉斯算子卷积[40], 提取高频残差 (high-frequency residual, HFR).

$$HFR = \sum_{C\in\{\mathrm{R,G,B}\}} X_C * K_{\mathrm{Lap}} \tag{2}$$

$$K_{\mathrm{Lap}} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{3}$$

其中, X_C 表示输入图像在第 C 个颜色通道上的二维像素矩阵, K_{Lap} 为 3×3 的 Laplacian 核, *表示二维卷积操作. 该操作在每个颜色通道上独立进行, 得到的高频残差伪影能够有效突出由扩散模型、GAN 等生成机制引入的局部不连续性与噪声模式. 经过该操作, 图像的大部分低频语义内容被过滤, 从而强制模型关注微观层面的纹理指纹. 为保持数值稳定性, 输入残差将进一步进行线性缩放后送入主干网络.

在特征提取阶段, 本分支采用基于残差网络的卷积主干结构对图像中的高频伪影进行分层建模. 首先, 输入高频残差首先通过一个 3×3 卷积层、批归一化和 ReLU 激活函数进行初步编码, 并通过最大池化操作进行空间下采样. 之后, 特征依次通过 3 个残差阶段 (layer1、layer2 和 layer3), 分别包含 3、4 和 6 个残差单元, 并在阶段之间通过步长为 2 的卷积实现空间下采样, 从而逐步扩大感受野并提升特征的语义表达能力, 用于实现对图像结构信息的逐步抽象表示, 最终得到输出空间特征图 F_{pixel}. 由于 F_{pixel} 直接源自高频残差信号, 其空间分布能够在一定程度上反映生成伪影的强度分布, 因此本文将其作为引导信号传递至块级分支. 为获得紧凑且全局一致的伪影表示, 本分支对特征图 F_{pixel} 施加自适应全局平均池化, 得到的向量聚合了整幅图像的像素级伪影统计信息, 为后续的判别与对比学习提供统一表征空间. 为了从混合伪影中进一步剥离出代表特定生成器的判别性指纹, 本文在全局平均池化之后设计了专用的投影头, 如果令池化后的特征向量为 $P = GAP(F_{\mathrm{pixel}})$, 则投影逻辑可以表示为:

$$f_s = \phi_{s2}(\sigma(\phi_{s1}(P))) \tag{4}$$

其中, ϕ_{s1} 和 ϕ_{s2} 为线性变化层, σ 为非线性激活函数. 该显式建模放大不同生成模型在生成过程中的稳定差异性, 同时抑制与具体内容和语义无关的共性干扰因素, 从而提升模型溯源任务的判别性与泛化能力, 随后对 f_s 归一化, 得到域相关特征 $\hat{f}_s$:

$$\hat{f}_s = \frac{f_s}{\|f_s\|_2 + \eta} \tag{5}$$

其中, η 为平滑项. 该过程实现了域相关解耦, 得到的特征独立于块级分支提取的语义表征, 在最终分类阶段通过级联操作与块级表征结合, 共同构建多尺度生成模型归因指纹.

2.4 块级生成模式感知分支

与细粒度像素级高频伪影建模分支侧重于捕获局部高频伪影不同, 块级生成模式感知分支旨在从补丁尺度 (patch-level)[14]建模生成图像在结构布局、区域一致性以及语义组织层面的潜在偏差. 已有研究表明, 扩散模型与 GAN 在生成过程中, 往往会在图像的中高层结构、跨区域依赖关系以及全局语义一致性方面留下具有区分性的模式差异, 而此类信息难以仅通过局部像素统计进行有效建模[41]. 基于此, 本文设计了一种融合 ViT 的块级表征分支, 并引入跨分支交叉注意力机制, 实现像素级伪影信息对块级语义建模过程的引导式融合.

在块级特征建模过程中, 首先设计块级层面的语义与残差解耦编码模块, 给定三通道 RGB 图像 X, 采用基于

卷积的补丁嵌入 (patch embedding) 模块将输入图像划分为一组互不重叠的局部图像块, 并映射初始补丁表示 P. 在传统图像处理与频域分析中, 低频成分通常指图像中变化平缓、空间连续的结构信息, 例如背景区域、整体颜色分布与物体轮廓等. 这类信息通常通过高斯滤波、平均池化或频域低通滤波等方式获得, 其本质是在空间上对局部邻域进行平滑聚合, 从而抑制高频细节与噪声[42]. 然而, 这类固定滤波方式仅依赖空间邻域, 缺乏对全局语义关系的建模能力, 容易导致语义边界模糊以及结构信息过度平滑.

不同于上述低频提取方式, 本文引入自注意力机制对补丁序列进行全局建模, 得到语义特征 S, 该过程通过建模补丁之间的长程依赖关系, 有助于捕获图像的整体结构布局与跨区域语义一致性信息. 与此同时, 引入注意力聚合操作对补丁表示进行特征整合. 具体地, 该过程使用频率注意力机制通过对补丁之间的相关性进行加权融合, 使得空间上具有一致性与连续性的区域被强化, 而局部突变与细粒度纹理信息则相对被抑制. 从表征角度来看, 该类由注意力机制聚合得到的表示在空间分布上呈现出与传统低频成分相似的平滑性与全局一致性, 因此可视为对低频成分的一种近似建模.

基于上述分析, 本文采用自注意力输出替代传统低频分量进行残差构造, 相较固定滤波获得的低频成分, 自注意力聚合通过自适应加权实现对特征的平滑聚合, 能够在保留结构信息的同时避免过度平滑, 从而使残差项更加集中于语义不一致区域与异常纹理模式. 在此基础上, 通过在补丁表示中减去上述低频成分, 利用残差形式显式构造块级高频伪影特征 H_p:

$$H_p = P\text{-}Attn_{\text{freq}}(P,P,P) \tag{6}$$

其中, 输入的 3 个 P 分别对应查询 (Query)、键 (Key) 和值 (Value), 由于采用自注意力机制, 三者均由初始补丁表示 P 构成.

通过该设计, 模型能够在补丁尺度上有效抑制低频平滑成分, 迫使网络关注扩散模型或 GAN 在局部区域中引入的细微结构噪声与生成偏差. 为了让两路特征更好地发挥各自的作用, 同时考虑到在不同图像和不同生成模型中, 像素伪影对语义建模的重要性并不一致, 本文使用门控自适应融合机制实现块内自适应融合, 将语义特征 S 与归一化后的高频残差 $H_{p,\text{norm}}$ 融合, 生成自适应的融合权重 G_b:

$$G_b = \sigma(W_g[S \parallel H_{p,\text{norm}}]) \tag{7}$$

其中, W_g 表示门控自适应融合模块中的可学习线性映射参数, 用于对拼接后的语义特征 S 与归一化高频残差 $H_{p,\text{norm}}$ 进行特征重加权. 通过门控单元生成自适应融合权重, 从而得到增强后的补丁特征 T_{vit}, 表示如下:

$$T_{\text{vit}} = S + G_b \odot H_{p,\text{norm}} \tag{8}$$

该机制能够根据补丁内容复杂度动态调整对高频伪影的关注程度, 确保在语义主导区域保持稳定建模, 而在伪影显著区域增强判别能力.

2.5 跨分支注意力引导融合模块

本文设计如图 3 所示的像素级-块级跨分支注意力引导融合模块作为跨分支引导结构. 将细粒度像素级高频伪影建模分支提取的 F_{pixel} 作为全局背景下的指纹锚点, 对 T_{vit} 进行进一步的引导修正和增强. 由于像素级分支输出的是空间特征图, 而块级分支处理的是序列化标记, 我们首先将 F_{pixel} 展平并投影至与块级特征一致的嵌入维度, 设像素级高频伪影特征为 $F_{\text{pixel}} \in \mathbb{R}^{B\times C\times H\times W}$, 其中 B 是批量大小, C 是通道数, H 和 W 分别是特征图的高和宽. 首先进行展平操作, 将空间特征映射为序列表示为 $F_{\text{seq}} \in \mathbb{R}^{B\times H\times W\times C}$, 随后通过线性投影层映射至与块级分支一致的嵌入维度 D=768, 从而得到像素级伪影序列表示:

$$T_{\text{res}} = F_{\text{seq}}W_p + b_p,\ T_{\text{res}} \in \mathbb{R}^{B\times H\times W\times D} \tag{9}$$

其中, $W_p \in \mathbb{R}^{C\times D}$ 表示线性投影矩阵, b_p 为偏置项.

为了确保指纹提取分支的稳定性, 在跨分支融合过程中对 T_{res} 进行梯度阻断 (detach) 操作, 从而实现单向的信息引导机制. 从梯度传播角度来看, 在未进行梯度阻断时, 块级分支的损失函数梯度将通过跨分支连接反向传播至像素级分支, 使得像素分支的参数更新同时受到两种优化目标的影响: 一方面需要建模高频伪影特征, 另一方面又

需适应块级分支的全局语义建模需求. 这种跨分支的梯度耦合容易导致优化目标冲突, 使得像素级分支提取的低层统计特征被高层语义信息所干扰, 从而削弱其对生成模型指纹的表征能力. 为此本文的 detach 操作显式阻断从块级分支到像素级分支的梯度回传路径, 从而将原本的双向梯度传播转变为单向信息传递. 该策略使得像素级分支仅受自身监督信号约束, 能够专注于稳定学习与生成过程相关的高频统计模式; 同时, 块级分支仍可利用该特征进行全局语义建模, 从而在不干扰底层特征学习的前提下, 实现跨层信息融合. 从特征学习的角度来看, 该单向引导机制本质上实现了不同层级特征的优化解耦, 像素级分支更侧重于捕获与生成过程相关的低层统计模式, 而块级分支则关注语义结构信息, 通过限制梯度耦合, 可以有效避免不同层级特征之间的过度干扰与表示混叠, 从而提升整体特征表达的稳定性与判别能力.

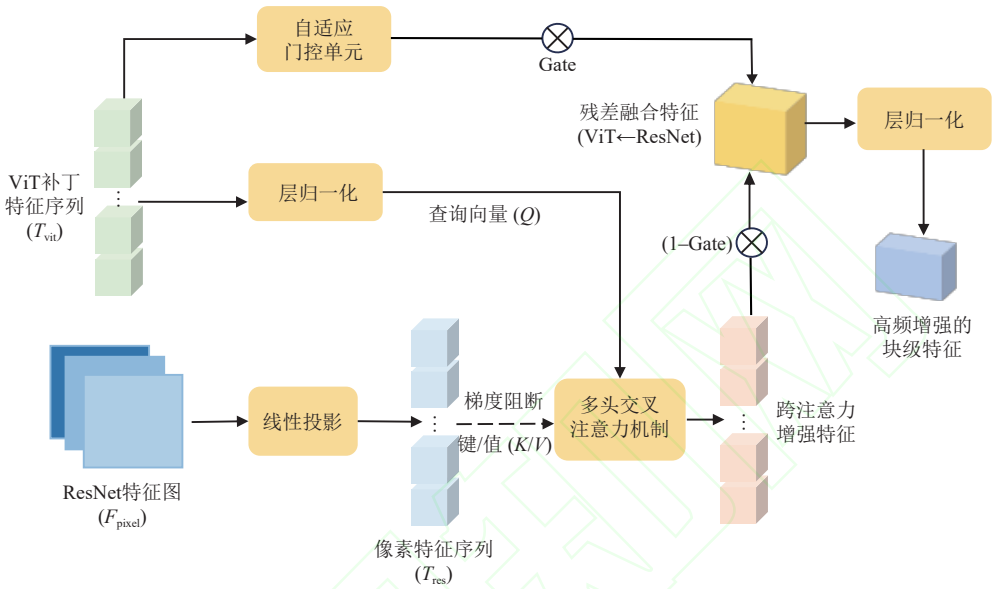


图 3 像素级-块级跨分支注意力引导融合模块结构框架图

在所设计的稳定的跨注意力模块中, 块级语义序列 T_{vit} 作为查询向量 (Query, Q), 像素级序列 T_{res} 作为键向量 (Key, K) 和值向量 (Value, V), 通过计算查询与键之间的相关性, 模型能够在全局语义建模过程中自动定位伪影信号最为显著的区域, 并将其信息注入语义表征之中, 通过交叉注意力机制实现跨分支信息交互, 上述过程可表示为:

$$Q = T_{vit}W_Q,\ K = T_{res}W_K,\ V = T_{res}W_V$$

$$T_{attn} = Softmax\left(\frac{QK^{T}}{\sqrt{d}}\right)V \tag{10}$$

其中, W_Q、W_K、W_V 为可学习投影矩阵, d 表示注意力头的维度.

为了进一步细化控制像素级信息对块级语义的影响程度, 本文在跨注意力输出后再次引入基于门控单元的融合策略. 该门控机制根据补丁语义特征动态生成融合权重, 实现对像素级伪影信息注入程度的自适应调节:

$$T_{fuse} = G_c \odot T_{attn} + (1 - G_c) \odot T_{vit} \tag{11}$$

其中, G_c 为由块级特征生成的门控权重. 最终, 通过门控残差连接与层归一化操作, 得到融合后的块级特征表示. 该设计确保了在伪影信号较弱的区域, 模型能够保持稳健的语义表征, 而在指纹特征显著的区域, 则能够有效增强其判别能力.

2.6 特征融合与归属预测

在前述两个分支中, 细粒度像素级高频伪影建模分支侧重于捕捉生成模型在低层统计空间中引入的局部指纹特征, 而块级生成模式感知分支则建模图像在结构与语义层面的全局生成规律. 为了充分利用两类互补信息, 本文

在归属预测阶段对不同分支的特征进行联合建模, 构建用于生成模型溯源的最终判别表示.

具体而言, 块级生成模式感知分支经过跨分支注意力引导融合与 Transformer 编码后, 输出一组融合后的补丁标记. 通过对该标记序列进行全局平均池化, 获得图像级的块级生成模式表示 f_b, 该向量刻画了生成模型在整体结构组织和语义布局上的一致性特征. 另一方面, 细粒度像素级高频伪影建模分支在高频残差信号的驱动下, 通过专用投影头提取出一个归一化的域相关伪影嵌入 f_p, 该向量主要编码不同生成模型在局部纹理、频谱分布等方面的差异性指纹. 而后本文采用特征级拼接的方式对二者进行联合表示建模:

$$f_{\text{cat}} = [f_b; f_p] \tag{12}$$

该联合表示可以同时保留全局生成结构信息与细粒度模型指纹信息, 为后续的归属判别提供更具判别性的特征. 随后, 融合特征通过全连接融合层进行映射, 并输入线性分类头以生成最终的生成模型归属预测分数. 通过上述特征聚合与判别建模方式, 模型能够在归属预测阶段同时利用生成模型在局部伪影层面和全局生成模式层面所留下的差异性信息, 从而在多模型复杂场景下显著提升生成图像溯源与拒识的鲁棒性和准确性.

2.7 联合优化目标函数

为实现对多源生成图像的有效归属判别, 并提升模型在未见生成模型场景下的鲁棒性, 本文在训练过程中设计了由闭集监督损失与伪影结构约束损失共同组成的联合优化目标函数. 同时, 在推理阶段引入基于分类输出能量评分的拒识机制, 以支持开放集条件下的未见模型检测.

模型首先通过主干网络提取融合特征表示 $f_{\text{fused},i}$ 并由线性分类器输出类别得分:

$$z_i = W f_{\text{fused},i} \tag{13}$$

其中, $W = [w_1, w_2, \ldots, w_C]^{\text{T}}$ 表示分类头权重矩阵, C 表示为已知生成模型类别数. 对于已知生成模型类别, 本文采用 *Softmax* 交叉熵损失对生成模型标签进行监督, 其定义为:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(w_{y_i}^{\text{T}} f_{\text{fused},i}\right)}{\sum_{j=1}^{C}\exp\left(w_j^{\text{T}} f_{\text{fused},i}\right)} \tag{14}$$

其中, B 表示批次大小, y_i 表示第 i 个样本的生成标签. 该损失用于引导模型学习不同生成模型之间的判别性表示, 是整体训练的主要监督信号, 确保了融合特征具有最基本的判别能力. 为进一步压缩同类生成模型在特征空间中的分布范围、增强类别间可分性, 本文在融合特征空间上引入中心损失 (center loss)[43]约束. 该损失通过约束样本特征向其对应类别中心聚拢, 有效减少类内方差, 其定义如下:

$$\mathcal{L}_{\text{center}} = \frac{1}{B}\sum_{i=1}^{B}\left\|f_{\text{fused},i} - c_{y_i}\right\|_2^2 \tag{15}$$

其中, C_{y_i} 表示类别 y_i 在融合特征空间中的中心向量. 该损失与交叉熵联合优化, 有助于降低类内分散度, 为后续开集判别提供更加稳定的特征基础. 同时, 为避免细粒度像素级高频伪影建模分支仅学习到相对相似性而缺乏明确的生成模型语义约束, 本文进一步引入伪未见辅助分类损失, 对伪影特征施加显式的生成模型判别监督. 具体而言, 我们将细粒度像素级高频伪影建模分支输出的特征通过独立的生成模型判别头进行投影, 并采用交叉熵损失引导模型区分不同生成模型来源, 从而在特征空间中建立稳定的类别锚点. 该辅助分类损失定义为:

$$\mathcal{L}_{\text{ascls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{C_g} y_i(g)\log\hat{y}_i(g) \tag{16}$$

其中, N 表示批次大小, C_g 为已知生成模型类别数, $y_i(g)$ 为第 i 个样本在生成模型类别 g 上的真实标签指示函数, $\hat{y}_i(g)$ 表示伪影判别头输出的 *Softmax* 预测概率, 该损失通过直接约束伪影特征与生成模型标签之间的对应关系, 有效防止对比学习过程中出现类别混淆问题, 并为后续的结构对齐提供判别基础.

在此基础上, 为了实现从复杂的图像内容中剥离出通用的生成指纹, 本文对伪影表征分支施加监督. 采用 NT-Xent (normalized temperature-scaled cross entropy)[44]形式的对比损失来增强同一生成模型样本在伪未见特征空间

中的结构一致性. 在批次内使同一生成源的样本构建正样本对, 表示为 $\left(z_i^{(k)}, z_j^{(k)}\right)$ 并将其余样本视为负样本, 迫使模型学习内容无关的生成模式, 基于此, 可以定义伪影特征对比学习损失为:

$$\mathcal{L}_{\text{ascon}} = -\frac{1}{2N}\sum_{k=1}^{N}\left[\log\frac{\exp\left(z_i^{(k)}\cdot z_j^{(k)}/\tau\right)}{\sum_{l\neq k}\exp\left(z_i^{(k)}\cdot z_l/\tau\right)} + \log\frac{\exp\left(z_i^{(k)}\cdot z_i^{(k)}/\tau\right)}{\sum_{l\neq k}\exp\left(z_j^{(k)}\cdot z_l/\tau\right)}\right] \tag{17}$$

其中, N 表示当前批次内构造的正样本对数量, z_l 表示当前批次中所有伪影特征向量, τ 表示温度系数, 用于缩放相似度的分布, 使对比学习更稳定. 最终损失对所有正样本对取平均, 并对 i 和 j 双向计算, 以增强特征空间的对称约束. 该损失促使同一生成模型的伪影特征在嵌入空间中聚集, 同时不同的生成模型伪影特征被有效拉开, 为开集拒识提供了可分辨的特征结构基础. 为防止补丁级高频残差特征在训练过程中出现数值波动, 引入如下频率正则化项:

$$\mathcal{L}_{\text{freq}} = \frac{1}{B}\sum_{i=1}^{B}\left(\left\|f_i^{\text{freq}}\right\|_2 - 1\right)^2 \tag{18}$$

其中, f_i^{freq} 表示第 i 个样本在块级生成模式感知分支中提取的伪影特征向量, 该特征是由经频率注意力聚合得到的补丁级高频残差. 该损失主要用于稳定块级高频特征分布, 避免训练过程中出现尺度不一致问题. 综合上述各项约束, 使用各权重系数 λ 平衡不同分支对整体训练目标的影响, 模型的总体优化目标可以定义为:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{center}}\mathcal{L}_{\text{center}} + \lambda_{\text{ascls}}\mathcal{L}_{\text{ascls}} + \lambda_{\text{ascon}}\mathcal{L}_{\text{ascon}} + \lambda_{\text{freq}}\mathcal{L}_{\text{freq}} \tag{19}$$

总的来说, 各损失项在优化过程中形成如下协同关系: 交叉熵与中心损失共同构建判别性特征空间, 辅助分类损失提供伪影特征的语义约束, 对比损失增强生成模式的一致性与可分性, 而频率正则项则确保特征分布的稳定性. 多项损失的协同优化, 使模型能够同时具备良好的判别能力、结构表达能力及泛化性能, 从而提升生成模型溯源任务的整体效果.

2.8 开集拒识评估机制

在推理阶段, 模型不仅需要对来自已知生成模型的样本进行准确归属判别, 还需对未在训练阶段出现的未见生成模型样本进行有效拒识. 为此, 本文首先引入基于能量的开集建模方法[45]替代最大 *Softmax* 概率 (maximum *Softmax* probability, MSP)[46]策略. 给定模型输出的类别得分, 从而获得更加稳定且可分的拒识评分.

给定输入样本 x, 模型输出分类的类别分数记为 $f(x) = [f_1(x), f_2(x), \ldots, f_C(x)]$, 其中 C 表示已知生成模型类别数. 其对应的能量评分函数定义为:

$$E(x;T) = -T\cdot\log\sum_{k=1}^{C}\exp\left(f_k(x)/T\right) \tag{20}$$

该能量函数刻画了样本在已知生成模型分布上的整体匹配程度. 在能量空间中, 已知类样本通常落入低能量区域; 而未见生成模型样本由于其伪影模式与训练分布不匹配, 往往无法在任何已知类别上形成显著响应, 表现为整体能量值偏高. 该策略可以避免 MSP 策略中过度依赖单一类别概率的问题, 在多生成模型复杂场景下具有更强的鲁棒性和稳定性.

为了全面评估模型在开集条件下对未见模型的拒识能力, 本文采用开放集分类率 (open-set classification rate, OSCR) 作为综合性能指标. OSCR 通过联合刻画正确分类率 (correct classification rate, *CCR*) 与未见样本的误接受率 (false positive rate, *FPR*) 之间的关系, 对模型性能进行整体评估[47]. 具体而言, 通过对判别阈值 δ 进行扫描, 在不同阈值条件下分别计算 *CCR* 与 *FPR*, 其定义如下:

$$CCR(\delta) = \frac{|\{x\in D_{\text{known}} \mid \hat{y}(x) = y(x), E(x;T) < \delta\}|}{|D_{\text{known}}|} \tag{21}$$

$$FPR(\delta) = \frac{|\{x\in D_{\text{unknown}} \mid E(x;T) < \delta\}|}{|D_{\text{unknown}}|} \tag{22}$$

其中, D_{known} 和 D_{unknown} 分别表示已知生成模型样本集合与未见生成模型样本集合, $\hat{y}(x)$ 和 $y(x)$ 分别为模型预测标签与真实标签. $E(x;T)$ 表示基于类别分数计算得到的能量评分, 该评分在已知类别样本上通常取较小值, 而在未见模型样本上呈现较大值. 通过对 *CCR-FPR* 曲线下的面积进行积分, 可得到 OSCR 指标值. 该指标能够综合反映

模型在保持已知生成模型高准确归属能力的同时对未见生成模型样本的抑制效果, 可以成为评估生成模型溯源任务中开集识别性能的有效度量.

3 实验结果与分析

3.1 实验设置

本节主要对实验中需要使用的数据集、训练方式以及配置细节做具体说明.

3.1.1 数据集与任务设置

为全面评估本文方法在生成图像归属识别与开放集场景下的鲁棒性与泛化能力, 本文搜集 CNN_synth_dataset[12]、GenImage[48]等多个经典论文数据集, 构建了包含真实图像和 18 种来自不同模型的生成图像的综合实验数据集. 该数据集包含 19 个类别, 共计 3 万张图像, 数据覆盖当前主流的生成范式, 包括 InfoMaxGAN、BEGAN、improved-diffusion 等多种 GAN 与 Diffusion 模型, 具有较强的代表性与多样性. 该数据集在本文中用作闭集模型溯源, 此数据集的模型图片如图 4 所示. 本文按照 8:1:1 的比例将完整数据集随机划分为训练集、验证集和测试集, 其中验证集用于模型参数与损失权重的调节, 测试集用于最终性能评估. 此外, 为模拟真实应用中未见生成模型的开放集场景, 本文从 GenImage 数据集中额外选取 WuKong、adm、sdv5、VQDM 这 4 种生成模型的数据, 在训练阶段完全不参与闭集模型溯源训练, 仅在测试阶段作为未见类别用于开放集拒识性能的验证. 在未见类别设置上, 每一类生成模型均包含约 1 000 张测试图像, 与已知类别的每类样本规模保持基本一致, 以避免由于样本数量差异过大对评估结果产生影响, 从而更客观地评估模型在未见类别上的判别能力.

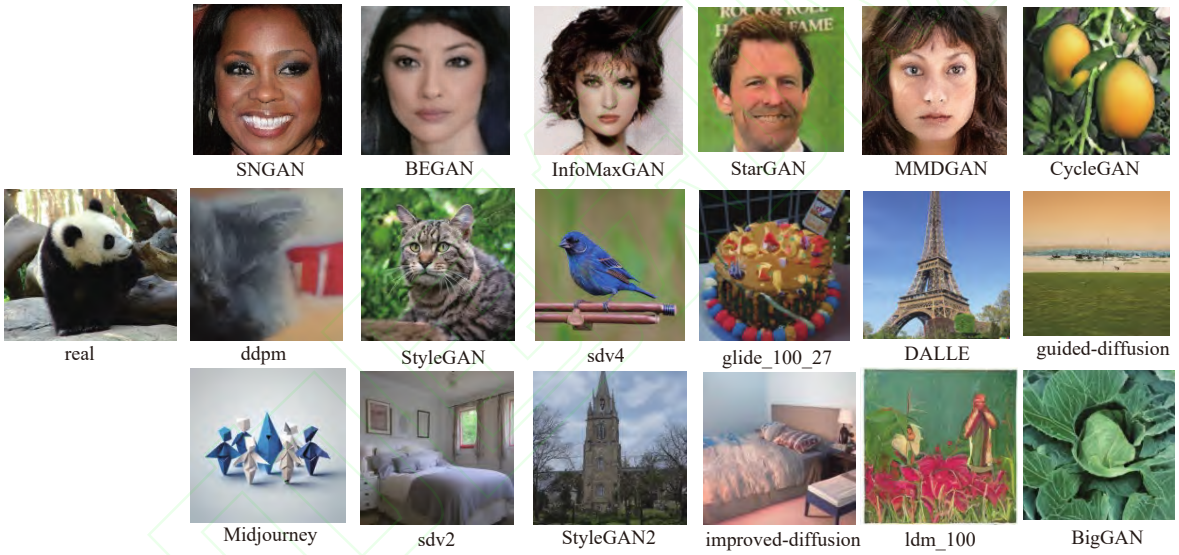


图 4 实验中闭集情况下所采用的模型图片示例

在任务设置上, 本文同时考虑闭集生成模型归属识别与开放集未见模型拒识两种评估场景. 在训练阶段, 仅使用上文所述 1 类真实和 18 类已知生成模型对应的图像样本进行监督学习; 在测试阶段, 除已知的 19 类生成模型样本作为闭集的测试集外, 进一步引入额外 4 类 (WuKong、adm、sdv5、VQDM) 未参与训练的生成模型作为未见类别, 以模拟真实应用中可能出现的未见生成模型情形. 对于闭集任务, 模型需要准确判断图像来源的具体生成模型, 而在开放集条件下, 模型不仅需要对已知生成模型进行正确归属, 还应能够有效区分并拒识未见生成模型样本. 上述设置使实验能够系统评估不同方法在已知多模型归属以及未见模型鲁棒性和泛化性等方面的综合性能.

3.1.2 实现细节与训练配置

在实验过程中, 输入图像在训练与测试阶段均统一随机裁剪为 224×224 的固定尺寸, 以保证不同数据源之间

的尺度一致性并提高模型训练的稳定性. 在模型训练过程中, 采用 Adam 优化器对网络参数进行更新, 训练轮数设置为 30 个训练轮次, 实验中批量大小 (batch size) 均设置为 16. 初始学习率设为固定值, 并采用分阶段衰减策略, 即每 5 个轮次对学习率进行一次衰减, 以平衡模型在前期的快速收敛与后期的精细优化. 该学习率调度策略有助于缓解训练后期的震荡现象, 并提升模型的最终泛化性能. 此外, 为减少随机性对实验结果的影响并提高可复现性, 本文在训练过程中固定随机种子 (seed=100), 并设置相关计算为确定性模式.

为充分挖掘不同生成模型之间的判别线索, 训练阶段联合优化多种损失函数, 包括用于生成模型归属预测的交叉熵损失、约束特征分布结构的中心损失以及针对生成伪影表示设计的对比学习损失. 各损失项通过加权方式进行联合优化, 其权重系数在验证集上进行调节确定. 所有实验均在配备 NVIDIA RTX 3090 GPU 的计算平台上完成, 平台环境为 PyTorch 1.13.1. 训练与测试过程中严格区分已知生成模型与未见生成模型数据, 以确保实验评估的公平性与可重复性. 本文中所报告的实验结果均基于相同的实现配置与训练策略获得.

3.1.3 评价指标

为全面评估模型在闭集生成模型归属与开放集未见生成模型识别场景下的性能, 本文采用多种评价指标进行综合衡量. 在闭集生成模型溯源任务中, 主要采用分类准确率 (accuracy, Acc) 作为评价指标, 用以衡量模型对已知生成模型来源的整体判别能力. 在开放集识别场景下, 本文重点关注模型区分已知生成模型样本与未见生成模型样本的能力. 为此, 采用 ROC 曲线下面积 (area under curve, AUC)[49]作为开集检测指标, 通过比较已知样本与未见样本的判别得分分布, 评估模型在不同阈值条件下的整体区分性能.

此外, 为进一步综合衡量模型在开放集条件下的实际应用能力, 本文引入 OSCR 作为核心评价指标. OSCR 同时考虑已知类别样本的正确分类率与对未见类别样本的拒识性能, 能够更全面地反映模型在真实未见生成模型场景中的鲁棒性与稳定性.

通过 Acc、AUC 与 OSCR 等指标的联合评估, 本文能够从闭集生成模型归属精度以及开放集未见生成模型区分与拒识能力等多个维度, 对所提方法进行全面分析.

3.2 方法对比实验结果与分析

为验证所提方法在生成图像溯源与开放场景识别任务中的有效性, 本文在包含真实图像和 18 类生成模型的数据集上, 将其与多种代表性方法在同一实验设置下进行系统对比. 对比方法涵盖传统闭集分类模型、基于伪影建模的溯源方法以及面向开放集场景的检测与归属方法, 具体包括 RepMix[4]、DNA-Det[50]、POSE[23]等多种代表性方法. 实验结果如表 1 所示, 对比结果分别从闭集分类平均准确率 (Acc)、AUC 以及 OSCR 这个指标进行评估.

表 1 本文方法 CPB-AM 与其他方法性能对比 (%)

方法	Acc	AUC	OSCR
SimpleCNN	56.74	48.53	30.69
ResNet50[51]	83.82	75.55	67.20
DNA-Det[50]	85.11	81.09	74.75
RepMix[4]	84.63	59.98	51.59
POSE[23]	78.42	83.16	72.51
NPR[13]	93.22	74.63	72.95
ReTD[26]	80.07	55.57	46.49
CLIDE[52]	63.73	66.40	30.60
AIDE[53]	91.35	75.91	72.81
CPB-AM (Ours)	**98.49**	**87.16**	**86.47**

注: 加粗数据表示最优结果

从表 1 可以观察到, 选取的基础卷积模型在该任务中存在明显性能瓶颈. 以 ResNet50 和 SimpleCNN 为代表的闭集分类模型在 Acc 指标上具有一定差距, 其中 ResNet50 模型在闭集上表现较好, 但二者在开集上 AUC 和 OSCR 均下降较多, 尤其是 SimpleCNN, 开集指标下降明显, 表明仅依赖判别式特征学习难以应对未见生成模型带

来的分布偏移问题. 这一现象说明, 传统 CNN 更倾向于学习内容相关特征, 而非稳定的生成指纹. 本文方法较基线的 ResNet50 方法在溯源准确率上提高了 14.67%, 说明本文的设计对多类别闭集处理的有效性.

针对生成模型溯源等相关任务设计的专用方法在开集性能上整体优于传统 CNN 模型. 具有代表性的 POSE 方法通过渐进式开放空间扩展机制, 在 AUC 和 OSCR 指标上取得了较为稳定的表现, 体现了其在开放集判别方面的优势. DNA-Det 与 ReTD 等方法主要基于生成伪影或重构残差进行建模, 在部分指标上表现稳定, 但可能受限于单一尺度视角, 整体性能表现上仍然存在一定局限. 关于 ReTD 方法, 在本文实验设置下的复现结果较其原文报道的闭集结果出现了明显下降, 分析认为可能是因为本文采用的数据集的模型类别数量较原文有扩展, 同时在保证统一训练策略的前提下, 单类样本规模相对原文有所减少, 而非实现层面的差异.

相较之下, 本文方法在 Acc 和 AUC 两项核心指标上均取得了最优结果, 相较在当前数据集上效果最好的方法分别提高了 5.27%、4.00%, 在 OSCR 指标上亦显著优于其他对比方法, 体现出所提方法在封闭集溯源与开放集拒识任务中的良好综合性能. 尽管数据集包含 18 类生成模型, 且不同模型在内容分布与生成机制上存在差异, 本文方法仍能够保持稳定优势, 验证了所提基于像素级-块级伪影协同建模策略在复杂生成模型分布下的鲁棒性与可扩展性.

进一步地, 为分析不同方法在具体生成模型类别上的识别差异, 表 2 给出各方法在真实图像及 18 类生成模型上的分类准确率对比结果. 从类别层面可以观察到, 传统闭集模型 (如 SimpleCNN 与 ResNet50) 在部分类别上表现波动明显. 一些基于伪影或残差建模的方法, 如 ReTD 等, 在特定类别上能够达到较高准确率, 但在另一些类别上存在明显下降. 这表明这些方法在跨模型泛化方面仍存在一定局限. 相较之下, 本文方法在绝大多数类别上均取得领先或接近最优的结果, 尤其是在真实图像及多种扩散模型上表现稳定, 同时在多类 GAN 模型上也保持较高识别准确率. 值得注意的是, 在模型数量较多、分布差异较大的复杂设置下, 本文方法未出现明显类别塌陷现象, 说明所提像素级-块级伪影协同策略能够在局部细粒度统计异常与上下文结构残差信息之间建立更稳定的判别边界.

表 2　本文方法 CPB-AM 与其他方法的各类别 Acc 对比 (%)

类别	SimpleCNN	ResNet50[51]	DNA-Det[50]	RepMix[4]	POSE[23]	NPR[13]	ReTD[26]	CLIDE[52]	AIDE[53]	CPB-AM (Ours)
real	6.67	48.00	11.58	56.00	78.67	89.33	56.00	83.67	77.33	**92.33**
glide-100-27	63.00	81.00	96.00	56.00	67.00	96.00	72.00	52.00	93.00	**98.00**
ddpm	76.00	89.00	85.00	88.00	75.33	92.00	73.00	48.70	92.67	**99.33**
InfoMaxGAN	94.50	**100.00**	**100.00**	97.00	**100.00**	**100.00**	98.50	40.50	98.00	**100.00**
sdv4	40.00	77.00	94.00	77.00	34.00	91.00	95.00	54.00	82.00	**96.00**
ldm_100	16.00	58.00	82.00	64.00	35.00	**98.00**	55.00	93.00	91.00	**98.00**
Midjourney	51.50	80.00	67.50	84.50	51.50	**94.00**	71.00	78.00	82.50	93.50
CycleGAN	37.88	74.24	**100.00**	87.88	80.30	98.48	66.67	50.00	87.88	**100.00**
sdv2	89.00	99.00	100.00	98.00	88.00	98.00	**100.00**	58.86	91.00	**100.00**
DALLE	66.00	86.50	78.50	77.00	98.00	91.00	98.00	92.00	90.00	**98.50**
MMDGAN	32.50	93.00	99.00	88.00	99.50	93.00	75.50	74.00	97.50	**100.00**
StyleGAN	54.33	84.00	63.33	95.67	92.33	**98.67**	71.33	54.33	93.67	98.00
StarGAN	87.06	98.01	**100.00**	98.51	99.50	**100.00**	99.00	87.06	99.00	**100.00**
StyleGAN2	83.67	94.33	90.67	98.00	89.33	98.67	88.33	69.67	97.67	**99.67**
BEGAN	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	46.50	**100.00**	71.00	99.50	**100.00**
BigGAN	19.00	75.50	87.00	82.00	78.50	98.50	75.50	52.00	95.50	**99.50**
SNGAN	43.00	91.50	98.50	86.00	97.00	92.50	88.50	56.50	89.50	**99.50**
guided-diffusion	41.50	68.50	65.00	83.00	53.50	95.50	63.00	55.50	82.00	**99.00**
improved-diffusion	76.50	95.00	99.00	91.50	72.50	**100.00**	75.00	40.00	96.00	**100.00**

注: 加粗数据表示最优结果

3.3　消融实验与分析

为了深入分析各关键设计模块对模型性能的影响, 本节从网络结构与机制设计、伪影建模与频域策略、多分支特征融合方式这 3 个方面进行消融实验, 并展示相应的性能变化.

3.3.1 结构与机制消融分析

为验证本文方法所添加模块的有效性, 本节以基本的 ResNet50 框架和 ViT 模型的双分支框架为基础框架, 依次引入像素级高频建模 (HF) 模块, 块级高频伪影建模 (PHF) 模块, 域相关解耦 (DA-Decoupling) 模块和跨分支注意力引导融合 (CBAGF) 机制, 用于构建逐步增强的生成指纹表示. 各模块配置及对应实验结果如表 3 所示.

表 3 各模块对模型性能的影响 (%)

模块	Acc	AUC	OSCR
基准框架	88.71	74.54	69.48
+HF	97.03	81.45	81.01
+HF+PHF	92.34	74.31	71.90
+HF+DA-Decoupling	94.82	76.00	74.30
+HF+PHF+DA-Decoupling	97.63	83.30	82.65
+HF+CBAGF+PHF+DA-Decoupling (CPB-AM)	**98.49**	**87.16**	**86.47**

注: 加粗数据表示最优结果

从表 3 数据可以观察到, 基准框架在未引入显式高频伪影建模的情况下, 尽管该策略在闭集分类任务中能够取得一定效果, 但在开放集场景下面对未见生成模型时, 其判别边界缺乏稳定的生成指纹约束, 导致 AUC 和 OSCR 指标存在明显性能瓶颈. 在基准框架上引入像素级高频特征建模 (HF) 后, 模型在闭集准确率、开集 AUC 及 OSCR 指标上均取得显著提升, 说明像素层面的高频伪影中蕴含与生成模型相关的稳定特征信息. 该结果验证了高频残差信息在生成模型溯源与未见类拒识任务中的重要作用, 同时也为后续更高层次的结构建模提供了可靠的特征基础. 然而, 在此基础上进一步引入块级高频伪影建模 (PHF) 时, 性能出现一定下降, 分析认为, 这是由于补丁级高频建模在缺乏约束情况下, 容易引入与图像内容和语义结构相关的干扰信息, 从而削弱模型对生成伪影的一致性建模能力, 进而影响判别稳定性. 在类似情况的实验中, 在仅引入 HF 模块的情况下加入域相关解耦 (DA-Decoupling) 模块, 虽然整体性能仍优于基准框架, 但相较于仅使用 HF 模块仍存在一定下降. 这表明, DA-Decoupling 模块本身需要与多源伪影特征协同作用, 才能充分发挥其抑制内容相关干扰的能力. 当 HF、PHF 和 DA-Decoupling 模块同时引入时, 模型性能得到明显恢复并进一步提升. 该结果可以体现出块级伪影特征在显式的域解耦约束下, 可以有效抑制语义噪声并突出生成模型相关差异, 从而发挥其在溯源与开放集识别任务中的优势.

在此基础上, 进一步引入跨分支注意力引导融合 (CBAGF) 机制后, 模型在闭集准确率、开集 AUC 和 OSCR 这 3 项核心指标上均取得最优性能. 该机制能够根据不同输入样本的特征分布, 自适应调节像素级伪影特征与块级结构特征的重要性权重, 在保证闭集判别性能的同时, 显著提升模型对未见生成模型的拒识能力. 综合来看, 各模块在不同层级上发挥了互补作用, 其协同设计有效构建了稳定且判别性强的多尺度生成指纹表示.

3.3.2 频域策略消融分析

为验证频域策略设计的合理性, 本文进一步对不同高频提取方式及邻域设置进行消融分析. 具体而言, 在像素级伪影建模阶段, 本文采用邻域为 4 的 Laplacian 算子作为高频提取方式, 并选取邻域为 8 的 Laplacian 算子以及另一种基于尺度变换的高频残差提取方法进行对比, 实验结果如表 4 所示.

表 4 不同频域策略对模型性能的影响 (%)

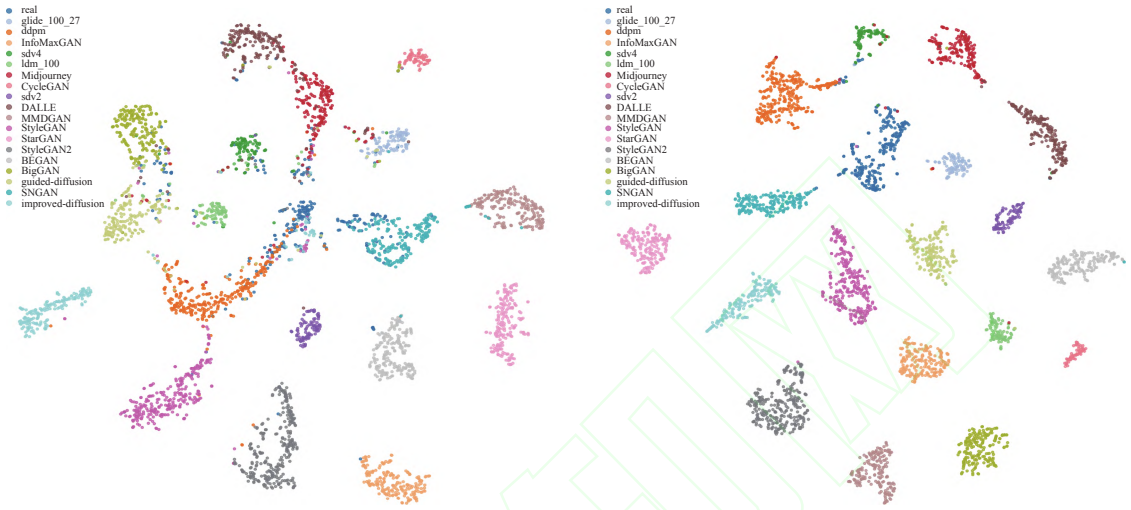
频域策略	Acc	AUC	OSCR
尺度变换	96.89	83.89	82.79
8邻域Laplacian算子	85.79	76.28	70.74
4邻域Laplacian算子	**98.49**	**87.16**	**86.47**

注: 加粗数据表示最优结果

从表 4 数据可以看出, 相较于本文所采用的 4 邻域 Laplacian 算子, 8 邻域 Laplacian 算子在闭集准确率上略有下降, 而在开放集 AUC 和 OSCR 指标上的性能下降更为明显. 这表明, 更大的邻域范围在增强边缘响应的同时, 也引入了更多与图像内容相关的噪声干扰, 从而削弱了模型在未见生成模型识别中的判别能力. 除此以外, 对像素

级高频特征的提取方法进行替换, 测试了一种基于尺度压缩与重建的高频残差近似提取方法, 即先对特征进行尺度变换, 再与原始特征作差以获得高频信息. 实验结果表明, 该方法同样能够提取一定程度的高频特征, 并取得相对较好的性能, 但在 3 项指标上均略低于 4 邻域 Laplacian 算子. 这说明直接在像素层面进行局部高频响应建模, 更有利于捕捉生成模型所引入的细粒度伪影模式.

为直观展示不同频域策略在特征分布上的差异, 本文对 4 邻域与 8 邻域 Laplacian 算子提取的特征进行了 t-SNE 可视化分析, 其结果如图 5 所示. 可以观察到, 4 邻域 Laplacian 算子所提取的特征在类别间具有更清晰的分布结构, 而 8 邻域 Laplacian 算子在特征分布上存在一定程度的混叠, 进一步验证了频域策略和邻域设置对开集识别性能的关键影响.



(a) 8邻域Laplacian算子 (30 epoch)　　(b) 4邻域Laplacian算子 (30 epoch)

图 5　不同邻域的 Laplacian 算子核训练闭集结果对比图

3.3.3　多分支特征融合方式消融分析

由于本文方法采用多分支结构并引入特征融合策略, 所以本节进一步分析融合不同特征对模型性能的影响. 为此, 本文对是否将未解耦的像素级分支特征直接拼接至最终表示中进行对比实验, 其配置及结果如表 5 所示.

表 5　表 5 不同特征融合方式对模型性能的影响 (%)

特征	Acc	AUC	OSCR
像素分支+解耦后的像素分支+块级分支	97.87	71.58	71.04
解耦后的像素分支+块级分支 (CPB-AM)	**98.49**	**87.16**	**86.47**

注: 加粗数据表示最优结果

从实验结果可以看出, 当同时拼接原始像素级分支特征、解耦后的像素级特征以及块级特征时, 模型在闭集分类任务中仍能保持较高准确率, 但在开放集 AUC 和 OSCR 指标上出现显著下降. 其中, OSCR 从 86.47% 降至 71.04%, 表明未解耦的像素级特征中包含较强的生成域相关信息, 其直接融合会干扰模型对未见类别的判别边界建模. 相比之下, 仅融合解耦后的像素级特征与块级特征能够显著提升模型在开放集场景下的检测性能. 这一结果表明, 通过显式的特征解耦与选择性融合, 模型能够在保留生成模型判别信息的同时, 有效抑制内容和语义相关干扰, 从而对未见生成模型实现有效拒识.

综合来看, 多分支特征融合并非是将所有特征直接融合的过程, 而需要在充分解耦与约束的前提下进行有选择的融合. 本文所采用的融合策略在闭集溯源性能与开放集拒识能力之间取得了良好的平衡, 进一步验证了 CPB-AM 框架设计的合理性.

4 总 结

本文围绕 AI 生成图像模型溯源任务中多尺度伪影难以协同建模、未见生成源场景下泛化能力不足的问题, 提出了一种基于像素到块级伪影协同建模 (CPB-AM) 的 AI 生成图像模型溯源框架. 该框架从生成图像伪影的多层次特性出发, 构建了像素级高频伪影建模与块级生成模式感知的双分支结构, 分别聚焦于局部纹理统计偏差与中高层结构语义及生成模式差异, 并通过引导式特征融合机制实现跨尺度信息协同, 充分挖掘不同尺度特征之间的互补性, 提升整体判别能力. 在包含 1 种真实图像和 18 种生成模型图像的数据集上的闭集实验结果表明, 所提方法在闭集生成模型溯源任务上展现出稳定的识别能力, 同时, 在开放集实验设置下, 所提框架在未见生成源识别与拒识任务中依然表现出较好的鲁棒性, 验证了像素到块级伪影协同建模策略在生成图像模型溯源任务中的有效性.

References

[1] Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, Aila T. Alias-free generative adversarial networks. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2021. 66.

[2] He ZJ, Li GB. Review of personalized image generation methods based on diffusion models. Ruan Jian Xue Bao/Journal of Software, 2026, 37(4): 1854–1884 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7511.htm [doi: 10.13328/j.cnki.jos.007511]

[3] Nightingale SJ, Farid H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proc. of the National Academy of Sciences of the United States of America, 2022, 119(8): e2120481119. [doi: 10.1073/pnas.2120481119]

[4] Bui T, Yu N, Collomosse J. RepMix: Representation mixing for robust attribution of synthesized images. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 146–163. [doi: 10.1007/978-3-031-19781-9_9]

[5] Karageorgiou D, Papadopoulos S, Kompatsiaris I, Gavves E. Any-resolution AI-generated image detection by spectral learning. In: Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE Computer Society, 2025. 18706–18717. [doi: 10.1109/CVPR52734.2025.01743]

[6] Tao RS, Tan CC, Liu H, Wang JK, Qin HT, Chang YK, Wang W, Ni RR, Zhao Y. SAGNet: Decoupling semantic-agnostic artifacts from limited training data for robust generalization in deepfake detection. IEEE Trans. on Information Forensics and Security, 2025, 20: 6429–6442. [doi: 10.1109/TIFS.2025.3581726]

[7] Yu N, Skripniuk V, Abdelnabi S, Fritz M. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE Computer Society, 2021. 14428–14437. [doi: 10.1109/ICCV48922.2021.01418]

[8] Wang ZT, Chen C, Zeng Y, Lyu LJ, Ma SQ. Where did I come from? Origin attribution of AI-generated images. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 3256.

[9] Wang ZT, Sehwag V, Chen C, Lyu LJ, Metaxas DN, Ma SQ. How to trace latent generative model generated images without artificial watermark? In: Proc. of the 41st Int'l Conf. on Machine Learning. Vienna: PMLR, 2024. 51396–51414.

[10] Sha ZY, Li Z, Yu N, Zhang Y. DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. In: Proc. of the 2023 ACM SIGSAC Conf. on Computer and Communications Security. Copenhagen: ACM, 2023. 3418–3432. [doi: 10.1145/3576915.3616588]

[11] Xu K, Zhang LZ, Shi JB. Detecting origin attribution for text-to-image diffusion models. In: Proc. of the 2025 IEEE/CVF Winter Conf. on Applications of Computer Vision. Tucson: IEEE Computer Society, 2025. 8775–8785. [doi: 10.1109/WACV61041.2025.00850]

[12] Wang SY, Wang O, Zhang R, Owens A, Efros AA. CNN-generated images are surprisingly easy to spot… for now. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2020. 8692–8701. [doi: 10.1109/CVPR42600.2020.00872]

[13] Tan CC, Liu H, Zhao Y, Wei SK, Gu GH, Liu P, Wei YC. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2024. 28130–28139. [doi: 10.1109/CVPR52733.2024.02657]

[14] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2021.

[15] Shi ZN, Chen HP, Zhang D, Shen XJ. Pre-training-driven multimodal boundary-aware vision Transformer. Ruan Jian Xue Bao/Journal of

Software, 2023, 34(5): 2051–2067 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6768.htm [doi: 10.13328/j.cnki.jos.006768]

[16] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[17] Chen CFR, Fan Q, Panda R. CrossViT: Cross-attention multi-scale vision Transformer for image classification. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE Computer Society, 2021. 347–356. [doi: 10.1109/ICCV48922.2021.00041]

[18] Chen FJ, Zhu F, Wu QX, Hao YM, Wang ED, Cui YG. A survey about image generation with generative adversarial nets. Chinese Journal of Computers, 2021, 44(2): 347–369 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2021.00347]

[19] Li ML, Qian ZX, Zhang XP. Survey on artificial intelligence-generated image detection. Journal of Image and Graphics, 2026, 31(1): 13–44 (in Chinese with English abstract). [doi: 10.11834/jig.250053]

[20] Xie TQ, Wu YY, Jing C, Sun WH. Survey of image detection methods generated by GAN models. Computer Engineering and Applications, 2024, 60(22): 74–86 (in Chinese with English abstract). [doi: 10.3778/j.issn.1002-8331.2405-0346]

[21] Luo YP, Du JL, Yan K, Ding SH. LaRE^2: Latent reconstruction error based method for diffusion-generated image detection. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2024. 17006–17015. [doi: 10.1109/CVPR52733.2024.01609]

[22] Yang YQ, Qian ZH, Zhu Y, Russakovsky O, Wu Y. D^3: Scaling up deepfake detection by learning from discrepancy. In: Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE Computer Society, 2025. 23850–23859. [doi: 10.1109/CVPR52734.2025.02221]

[23] Yang TY, Wang DD, Tang F, Zhao XY, Cao J, Tang S. Progressive open space expansion for open-set model attribution. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE Computer Society, 2023. 15856–15865. [doi: 10.1109/CVPR52729.2023.01522]

[24] Girish S, Suri S, Rambhatla S, Shrivastava A. Towards discovery and attribution of open-world GAN generated images. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE Computer Society, 2021. 14074–14083. [doi: 10.1109/ICCV48922.2021.01383]

[25] Pavlichenko N, Ustalov D. Best prompts for text-to-image models and how to find them. In: Proc. of the 46th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Taipei: ACM, 2023. 2067–2071. [doi: 10.1145/3539618.3592000]

[26] Chen WZ, Yuan FF, Cao C, Peng K, Wang DK, Liu YB. ReTD: Reconstruction-based traceability detection for generated images. In: Proc. of the 2025 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Hyderabad: IEEE, 2025. 1–5. [doi: 10.1109/ICASSP49660.2025.10890492]

[27] Qian YY, Yin GJ, Sheng L, Chen ZX, Shao J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer Int'l Publishing, 2020. 86–103. [doi: 10.1007/978-3-030-58610-2_6]

[28] Shiohara K, Yamasaki T. Detecting deepfakes with self-blended images. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE Computer Society, 2022. 18699–18708. [doi: 10.1109/CVPR52688.2022.01816]

[29] Almestekawy A, Zayed HH, Taha A. Deepfake detection: Enhancing performance with spatiotemporal texture and deep learning feature fusion. Egyptian Informatics Journal, 2024, 27: 100535. [doi: 10.1016/j.eij.2024.100535]

[30] Yan ZY, Zhang Y, Fan YB, Wu BY. UCF: Uncovering common features for generalizable deepfake detection. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE Computer Society, 2023. 22355–22366. [doi: 10.1109/ICCV51070.2023.02048]

[31] Koutlis C, Papadopoulos S. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 394–411. [doi: 10.1007/978-3-031-73220-1_23]

[32] Lee H, Yi J. Few-shot class-incremental model attribution using learnable representation from CLIP-ViT features. arXiv:2503.08148, 2025.

[33] Song HJ, Khayatkhoei M, AbdAlmageed W. ManiFPT: Defining and analyzing fingerprints of generative models. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2024. 10971–10981. [doi: 10.1109/CVPR52733.2024.01026]

[34] Yang TY, Wang DD, Cao J, Xu C. Model synthesis for zero-shot model attribution. IEEE Trans. on Multimedia, 2025, 27: 8967–8980. [doi: 10.1109/TMM.2025.3607778]

[35] Nguyen TD, Azizpour A, Stamm MC. Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of AI-generated images. In: Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE Computer Society, 2025. 3040–3050. [doi: 10.1109/CVPR52734.2025.00289]

[36] Sun ZM, Shen C, Yao TP, Yin BJ, Yi R, Ding SH, Ma LZ. Contrastive pseudo learning for open-world deepfake attribution. In: Proc. of

the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE Computer Society, 2023. 20825–20835. [doi: 10.1109/ICCV51070.2023.01909]

[37] Sharma S, Rani S, Dogra A, Bhatt MW. Edge-aware multisensor brain image fusion via guided filtering in Laplacian domain. Discover Artificial Intelligence, 2025, 5: 162. [doi: 10.1007/s44163-025-00446-y]

[38] Wang HH, Wu XD, Huang ZY, Xing EP. High-frequency component helps explain the generalization of convolutional neural networks. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2020. 8681–8691. [doi: 10.1109/CVPR42600.2020.00871]

[39] Odena A, Dumoulin V, Olah C. Deconvolution and checkerboard artifacts. Distill, 2016, 1(10): e3. [doi: 10.23915/distill.00003]

[40] Paris S, Hasinoff SW, Kautz J. Local Laplacian filters: Edge-aware image processing with a Laplacian pyramid. ACM Trans. on Graphics (TOG), 2011, 30(4): 68. [doi: 10.1145/2010324.1964963]

[41] Corvi R, Cozzolino D, Poggi G, Nagano K, Verdoliva L. Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE Computer Society, 2023. 973–982. [doi: 10.1109/CVPRW59228.2023.00104]

[42] Rahaman N, Baratin A, Arpit D, Draxler F, Lin M, Hamprecht F, Bengio Y, Courville A. On the spectral bias of neural networks. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 5301–5310.

[43] Wen YD, Zhang KP, Li ZF, Qiao Y. A discriminative feature learning approach for deep face recognition. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 499–515. [doi: 10.1007/978-3-319-46478-7_31]

[44] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proc. of the 37th Int'l Conf. on Machine Learning. Vienna: PMLR, 2020. 1597–1607.

[45] Liu WT, Wang XY, Owens JD, Li YX. Energy-based out-of-distribution detection. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1802.

[46] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[47] Dhamija AR, Günther M, Boult TE. Reducing network agnostophobia. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 9175–9186.

[48] Zhu MJ, Chen HT, Yan QY, Huang XD, Lin GY, Li W, Tu ZJ, Hu HL, Hu J, Wang YH. GenImage: A million-scale benchmark for detecting AI-generated image. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 3398.

[49] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8): 861–874. [doi: 10.1016/j.patrec.2005.10.010]

[50] Yang TY, Huang ZY, Cao J, Li L, Li XR. Deepfake network architecture attribution. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2022. 4662–4670. [doi: 10.1609/aaai.v36i4.20391]

[51] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[52] Betser R, Hofman O, Vainshtein R, Gilboa G. General and domain-specific zero-shot detection of generated images via conditional likelihood. In: Proc. of the 2026 IEEE/CVF Winter Conf. on Applications of Computer Vision. Tucson: IEEE Computer Society, 2026. 7809–7820. [doi: 10.1109/WACV61042.2026.00754]

[53] Yan SL, Li OX, Cai JY, Hao YB, Jiang XL, Hu Y, Xie WD. A sanity check for AI-generated image detection. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025.

附中文参考文献

[2] 何子健, 李冠彬. 基于扩散模型的个性化图像生成方法综述. 软件学报, 2026, 37(4): 1854–1884. http://www.jos.org.cn/1000-9825/7511.htm [doi: 10.13328/j.cnki.jos.007511]

[15] 石泽男, 陈海鹏, 张冬, 申铉京. 预训练驱动的多模态边界感知视觉 Transformer. 软件学报, 2023, 34(5): 2051–2067. http://www.jos.org.cn/1000-9825/6768.htm [doi: 10.13328/j.cnki.jos.006768]

[18] 陈佛计, 朱枫, 吴清潇, 郝颖明, 王恩德, 崔芸阁. 生成对抗网络及其在图像生成中的应用研究综述. 计算机学报, 2021, 44(2): 347–369. [doi: 10.11897/SP.J.1016.2021.00347]

[19] 李美玲, 钱振兴, 张新鹏. 人工智能生成图像检测技术综述. 中国图象图形学报, 2026, 31(1): 13–44. [doi: 10.11834/jig.250053]

[20] 谢天圻, 吴媛媛, 敬超, 孙伟恒. GAN 模型生成图像检测方法综述. 计算机工程与应用, 2024, 60(22): 74–86. [doi: 10.3778/j.issn.1002-8331.2405-0346]

作者简介

范宇祺, 硕士生, 主要研究领域为 AI 生成图像溯源和检测.

陈嘉欣, 博士, 讲师, 主要研究领域为多媒体内容安全, 人工智能安全, AI 生成内容检测.

廖鑫, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为人工智能安全, 信息内容安全.

陈昌盛, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为多媒体取证与内容安全, 文档图像篡改与翻拍检测, 物理层认证与二维码安全技术, 人脸活体检测与对抗攻防.

章登勇, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为多媒体信息安全, 图像处理与模式识别, 图像取证.